MINING OF TOP K DOMINATING QUERIES BASED ON PRIORITY OVER INCOMPLETE DATA

[1]**Sharanya Mohanan. C.V**, [2]**Magniya Davis**
[1]Student, [2]Assistant Professor
[1]Computer Science,
[1]St.Joseph's College, Thrissur, India

Abstract :* *Incompleteness of data is a serious issue and can have a significant effect on the conclusions that can be drawn from the data. Missing data can occur because of non response no information is provided for one or more items or whole unit. Incomplete data can also exists in wide spectrum of real datasets due to privacy preservation deice failure data loss and so on. In this paper for the first time we propose a method to handle incomplete data which consist of some missing dimensional values. Although some previous works are done but in this paper for the first time we introduce a new concept called priority value and this value is used to find top k dominating query over incomplete data[1]. The top k dominating query returns the K objects that dominates the maximum no of objects in a given dataset. Top K queries are also used for sliding window data streams[2].We formalize the problem and propose a skyline based algorithm to find top k objects in the result which dominates other objects based on dominance relationship.

IndexTerms – incomplete data, tkd query, priority assignment, local skyline, dominance relation.

I. INTRODUCTION

Data mining is the process of sorting through large datasets to identify patterns and establish relationships to solve problems through data analysis. Data mining allow enterprises to predict future trends. Real world data tends to be incomplete, noisy and inconsistent and an important task when preprocessing the data is to fill in missing values, smooth out noise and correct inconsistencies. Missing data is a serious issue and can have a significant effect on the conclusions that can be drawn from the data The TKD query returns the k objects that dominate the maximum no of objects in a given dataset. It combines the advantages of skyline and top k queries and plays an important role in many decision support applications. Missing data is defined as the data which is not stored for a particular variable in the observation of interest. The problem of missing data is relatively common n almost all research and handling of missing data is of significance. Given a set of S of d dimensional objects, the TKD query ranks the objects o in S dominated by o and returns the K objects from S that dominate the maximum no of objects. TKD query handles missing data based on priority value and a skyline based algorithm. A priority value is assigned to each dimensions of objects in dataset. This priority of the dimensions is also considered while finding the top elements. A skyline based algorithm is used here to find the resulting objects that are dominating the rest of the objects.

II. EXISTING SYSTEM

In this paper, we consider an incomplete dataset where some objects face he missing of attribute values in some dimensions, and study the problem of TKD query processing[1] over incomplete data. In particular, a TKD query on incomplete data returns the k objects that dominate the maximum no of objects from a given incomplete data set. An intuitive method for supporting the tkd query on incomplete data is to conduct pair wise comparisons among the whole dataset to get score of every object o that means the no of objects dominated by o and to return the k objects with highest scores. Priority values are not given. Approach is inefficient due to large size of dataset and the expensive cost of brute force based score computation. Data incompleteness is universal and querying incomplete data has become more and more important recently. It also triggered lots of effort in database community. Although the TKD query over incomplete data or uncertain data has been well studied ,TKD query processing over incomplete data still remains as a big challenge. This is because existing techniques cannot be applied to handle the TKD query over incomplete data efficiently.Another query similar to tkd is probabilistic tkd query over uncertain database retrieve uncertain objects in database that dominates other[3][4].

III. PROPOSED SYSTEM

In our proposed system we introduce a new concept called priority concept where priorities are assigned to each dimensions of objects in dataset. This priority value is also considered while finding the top elements in the given incomplete dataset. A skyline based algorithm is also introduced to find the top resulting objects that dominate the rest of the objects. Priority values and incomplete dataset is given as Input and the top k dominating objects will be the output. Priority values are given hence reduced time and increase efficiency. A data object is treated as an incomplete data object with missing value. Uncertainity of missing data is expressed in terms of some probability distributions. In our paper the priority concept helps the user to retrieve top k dominating objects from the rest of the objects based on priority values given by the user. One common way to find top k objects is that scoring all the objects with a scoring function[5].

IV. SYSTEM ARCHITECTURE

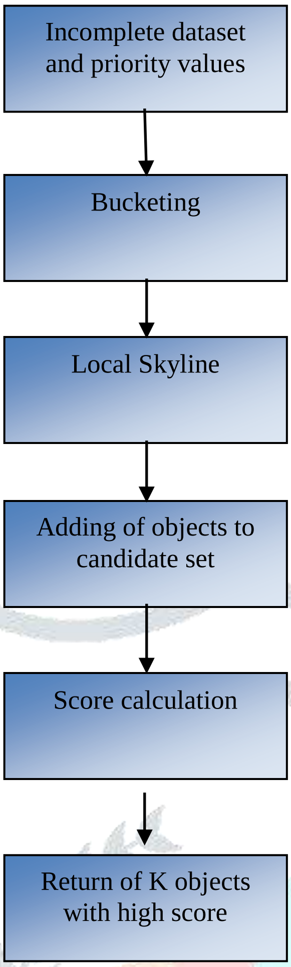

Incomplete dataset and priority values
Bucketing
Local Skyline
Adding of objects to candidate set
Score calculation
Return of K objects with high score

Fig 1.1:system architecture

In our system incomplete dataset and priority values are given as input. The output will be the k objects that will be considered as the top k objects that dominate the resulting objects. The system architecture in fig1.1 describes how the top k dominating objects are retrieved using priority concept and a skyline based algorithm. The dataset goes through various process. Incomplete dataset including some missing values and priority values are given as input then a process called bucketing is performed it is to cluster the objects with same bit vectors thus objects with same bit vectors will be inserted in a bucket then local skyline(top object) in the bucket is found and the score of ach objects in the candidate set with respect to entire dataset are calculated and objects with high scores are returned.

V. METHODOLOGY

This paper which introduces top k dominating query over incomplete data returns top k objects that dominates rest of the objects. The given dataset is incomplete that is, some dimension values of certain objects are missing due to various reasons. Some previous works are done on the topic, but in this project a new concept of priority value is introduced. It means a priority value is assigned to each dimensions of objects in dataset. This priority of the dimensions is also considered while finding the top elements. A skyline based algorithm is used here to find the resulting objects that are dominating the rest of the objects.

Fig1.2: an example of incomplete dataset

VEHICLE'S NAME	RATINGS BY USER TO EACH FEATURE				
	R1	R2	R3	R4	R5
Suzuki	_	_	3	2	1
Swift	3	1	_	_	2
Honda	1	_	_	2	1
Tata	2	_	3	_	7
Fiat	_	_	5	4	3

1.Dataset and priority values

Incomplete dataset and priority values are given as input and top k dominating objects are returned as output.

2.Bucketing

First clusters the objects in to buckets based on their bit vectors. It is calculated by, if object A(1,-,2,3) its bit vector will be 1011. That is, the missing value will be 0 and others will be 1. Objects with same bit vector will be inserted into same bucket.

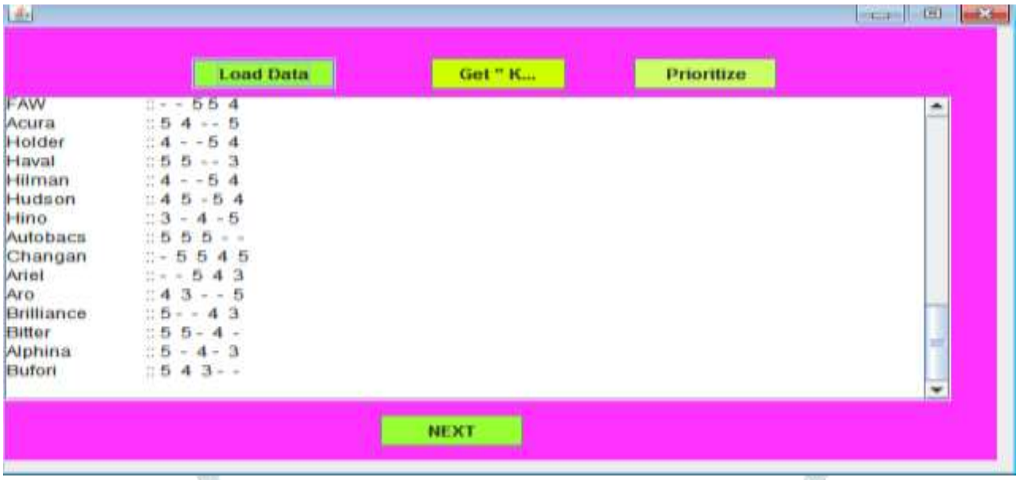
3.Local Skyline

Local Skyline is the top object in the bucket. Since each bucket contain the same bit vector local skyline of each bucket is found separately. For finding the local skyline the score of each element with respect to other are calculated. Then k elements with high score are added to a candidate set which is empty initially.

4.Score calculation

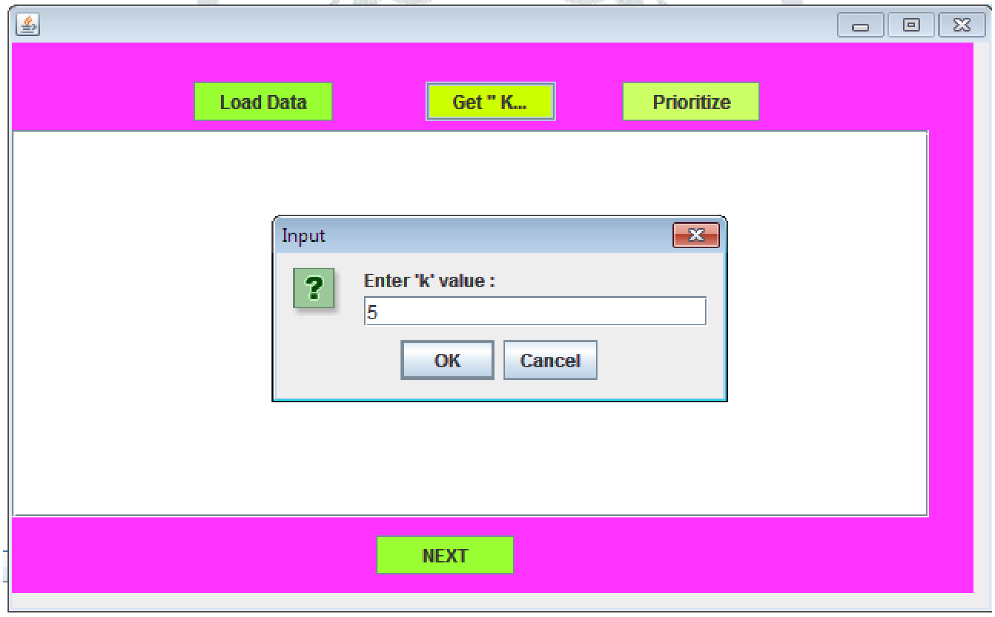
Score of an object a is the number of objects that A dominates.The dominance relationship can be defined as follows. The object A dominates B if and only if for all dimension i, Either a[i]<b[j] or one of them is missing. In this project since we are including the priority value for each dimension that is also considered. That is, in some cases where priority cannot be determined by the above definition. For example at one dimension. A dominats b and at second dimension. B dominates A then we cannot determine the dominance. In such cases object which dominate at the high priority dimension will selected as the dominating object. If they have same value in the high priority dimension then we will move to second priority object. Score of objects in candidate set . After finding the local skylines, they will be added to a candidate set. Then score of elements in the candidate set with respect to the entire dataset are calculated based on the score calculation technique discussed above. Then the K objects with high score from the candidate set are returned.

VI IMPLEMENTATION



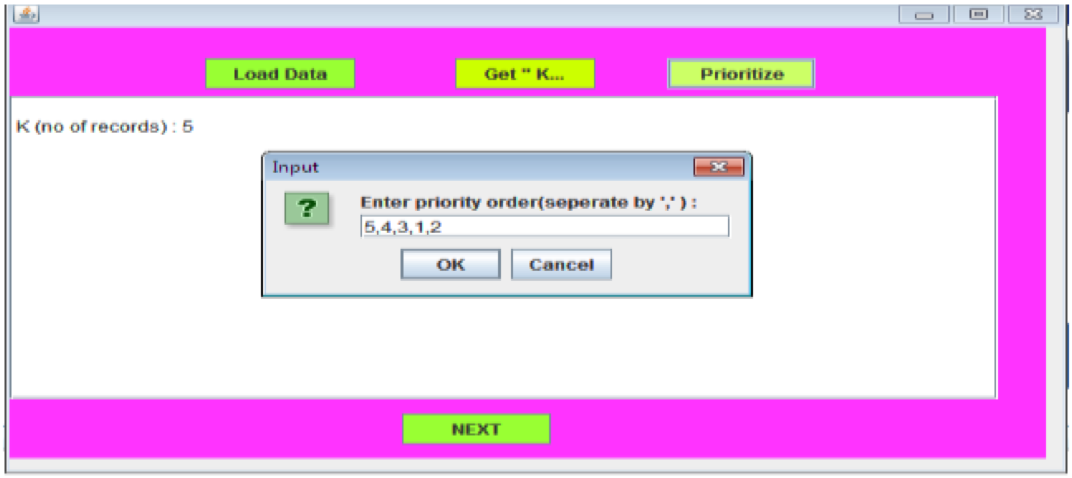
```
FAW         ::- - 5 5 4
Acura       ::5 4 - - 5
Holder      ::4 - - 5 4
Haval       ::5 5 - - 3
Hilman      ::4 - - 5 4
Hudson      ::4 5 - 5 4
Hino        ::3 - 4 - 5
Autobacs    ::5 5 5 - -
Changan     ::- 5 5 4 5
Ariel       ::- - 5 4 3
Aro         ::4 3 - - 5
Brilliance  ::5 - - 4 3
Bitter      ::5 5 - 4 -
Alphina     ::5 - 4 - 3
Bufori      ::5 4 3 - -
```

Fig1.3: load data



Enter 'k' value : 5

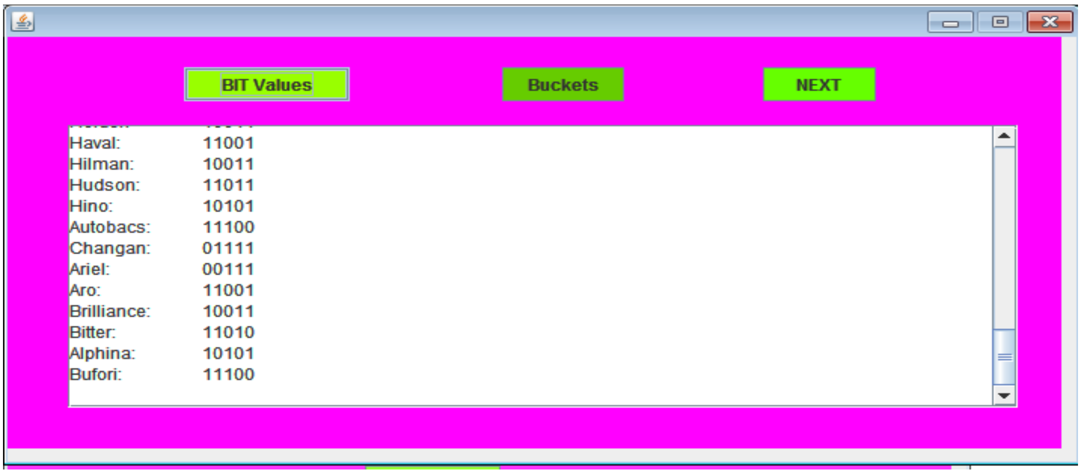
Fig1.4: entering k value



K (no of records) : 5

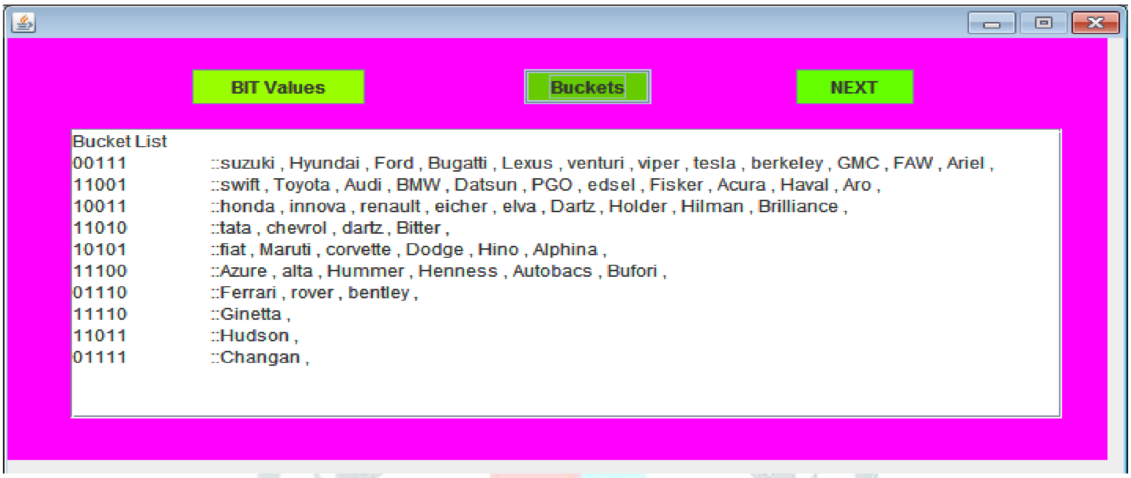
Enter priority order(seperate by ',') : 5,4,3,1,2

Fig1.5: entering priority



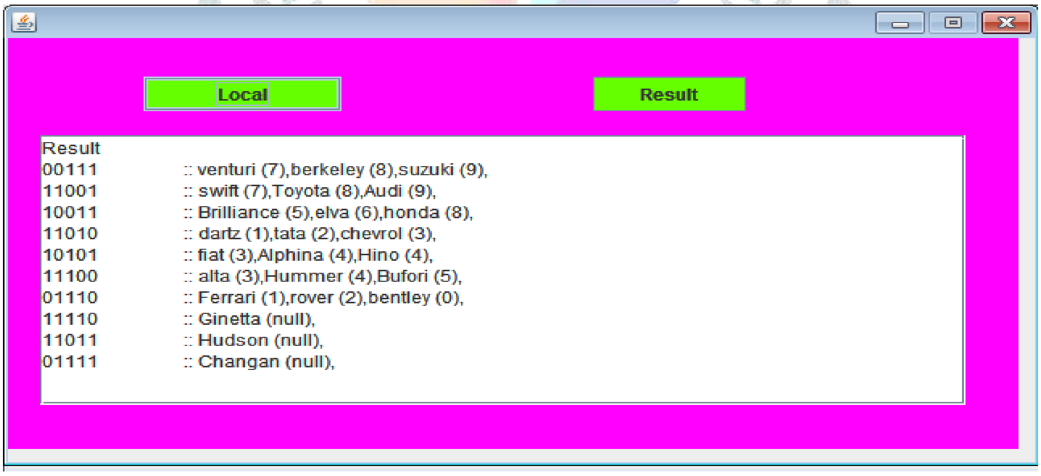
```
Haval:          11001
Hilman:         10011
Hudson:         11011
Hino:           10101
Autobacs:       11100
Changan:        01111
Ariel:          00111
Aro:            11001
Brilliance:     10011
Bitter:         11010
Alphina:        10101
Bufori:         11100
```

Fig1.6:converting to bit values



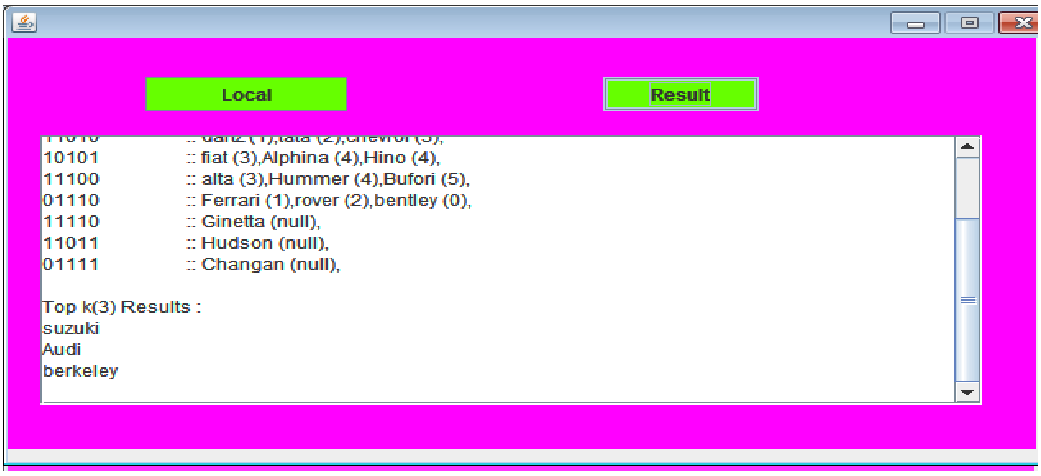
```
Bucket List
00111           ::suzuki , Hyundai , Ford , Bugatti , Lexus , venturi , viper , tesla , berkeley , GMC , FAW , Ariel ,
11001           ::swift , Toyota , Audi , BMW , Datsun , PGO , edsel , Fisker , Acura , Haval , Aro ,
10011           ::honda , innova , renault , eicher , elva , Dartz , Holder , Hilman , Brilliance ,
11010           ::tata , chevrol , dartz , Bitter ,
10101           ::fiat , Maruti , corvette , Dodge , Hino , Alphina ,
11100           ::Azure , alta , Hummer , Henness , Autobacs , Bufori ,
01110           ::Ferrari , rover , bentley ,
11110           ::Ginetta ,
11011           ::Hudson ,
01111           ::Changan ,
```

Fig1.7:bucketing



```
Result
00111           :: venturi (7),berkeley (8),suzuki (9),
11001           :: swift (7),Toyota (8),Audi (9),
10011           :: Brilliance (5),elva (6),honda (8),
11010           :: dartz (1),tata (2),chevrol (3),
10101           :: fiat (3),Alphina (4),Hino (4),
11100           :: alta (3),Hummer (4),Bufori (5),
01110           :: Ferrari (1),rover (2),bentley (0),
11110           :: Ginetta (null),
11011           :: Hudson (null),
01111           :: Changan (null),
```

Fig1.8:retrieving top elements from candidate set



```
10101           :: fiat (3),Alphina (4),Hino (4),
11100           :: alta (3),Hummer (4),Bufori (5),
01110           :: Ferrari (1),rover (2),bentley (0),
11110           :: Ginetta (null),
11011           :: Hudson (null),
01111           :: Changan (null),

Top k(3) Results :
suzuki
Audi
berkeley
```

Fig1.9: mined top k elements

VII CONCLUSION

Data mining is the process of discovering patterns in large datasets involvimg methods at the intersection of machine learning, statistics and database systems. It is an essential process to extract information from a dataset and transform it to an understandable form for further use. Data mining is primarily used today in companies with a strong consumer focus retail, financial, communication and marketing organizations to drill down into their transactional data and determine pricing, customer preferences and product positioning, impact on sales, customer satisfaction and corporate profits. Data mining becomes more complicated when missing values arise in a dataset and it can cause wrong results. Incompleteness is a serious issue and will have a significant effect on the conclusions drawn from the data. In this paper we demonstrate that how to mine top k elements from an incomplete dataset and also introducing a new concept priority assignment where priorities values are assigned and top k elements from the incomplete dataset are derived using the priority values and a skyline based algorithm.

VIII ACKNOWLEDGEMENT

First of all I am grateful to the Almighty GOD for establishing me to complete this project. I am especially thankful to my guide Ms. Magniya Davis, and all other faculty members from the department of computer science and my friends, for giving me their sole corporation and encouragement and critical I puts in the preparation of this report. Finally I express my gratitude to our lab instructors, colleagues, friends and my dear parents for giving valuable advice and support throughout my project.

REFERENCES

[1] Xiaoye Miao, Yunjun Gao, Baihua Zheng, Gang Chen, ?Huiyong Cui,"Top K dominating queries on incomplete data",in IEEE transactions on knowledge and data engineering on Data Engineering.

[2] Parisa Haghani,Sebastin Michael,Karl Aberer,"Evaluvating Top k Queries over incomplete data streams".

[3] X.Lian and L.chen,"Top k dominating queries in uncertain databases", in Proc.12th Int. Conference. Extending Database Technol : Adv. Database Technol,2009,pp.660-671.

[4] X.Lian and L.chen,"Probabilistic top k dominating queries in uncertain databases,"Inf.Sci.,vol 226,pp.23-46-2013.

[5] HabF,Yilas,George Beskales,Mohamad A. Soliman,"A survey on top k query processing on relational databases.